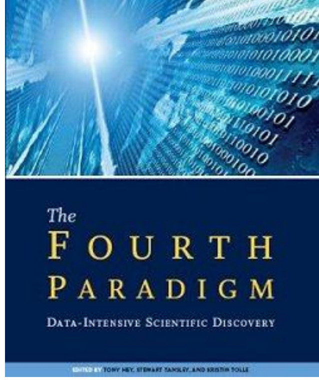
Life in the Cloud: TeraPixel and MODISAzure

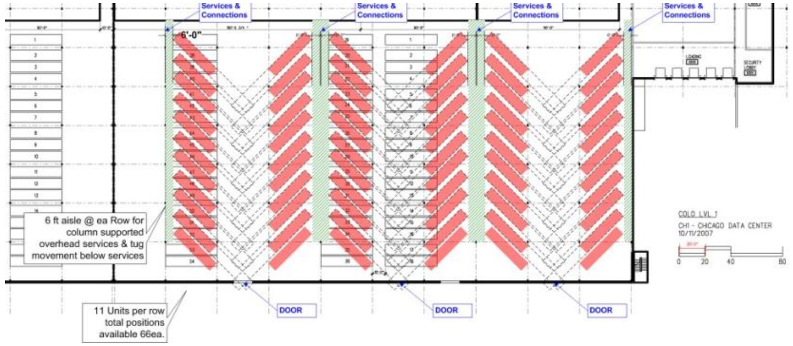
Catharine van Ingen
Partner Architect
eScience Group, Microsoft Research
CSIRO 26 October 2010

<http://research.microsoft.com/fourthparadigm>



Introduction

- The data deluge/landslide/tsunami/explosion is here
  - Science happens when multiple KB to PB datasets can be mashed up simply
- Commodity computing is here
  - Massive data centers, multi-core workstations, TB disks
- Yet resource, tedium, and complexity barriers exist
  - What programming paradigms are most efficient when people are the most significant cost?

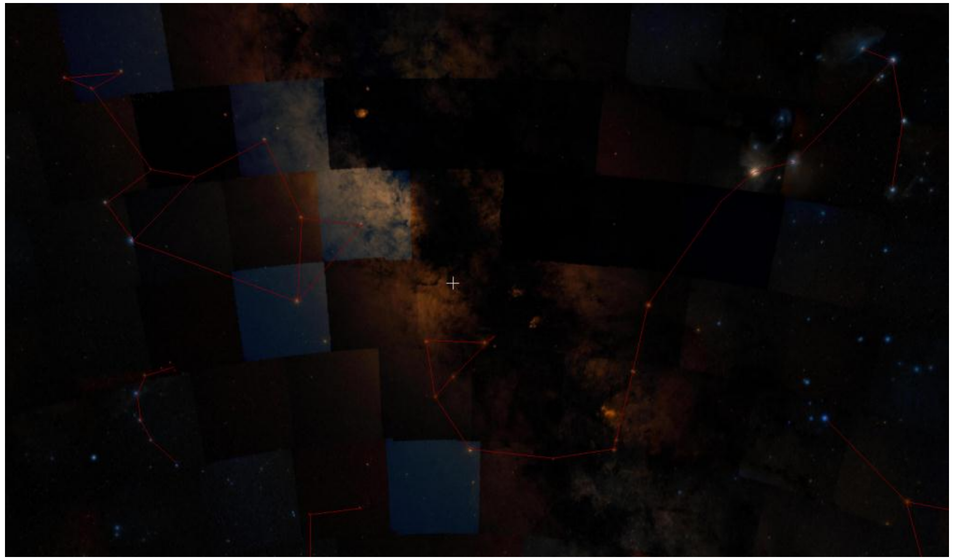


TeraPixel

I can see clearly, the rain is gone,
I can see all obstacles in my way.
Jimmy Clift

Seamless Visualization of the Night Sky

- Source imagery taken over 50 years by Palomar and Schmidt astronomical surveys
- 1791 image pairs (23040x23040 or 14000x14000 images) (4TB)
- Image process introduces artifacts



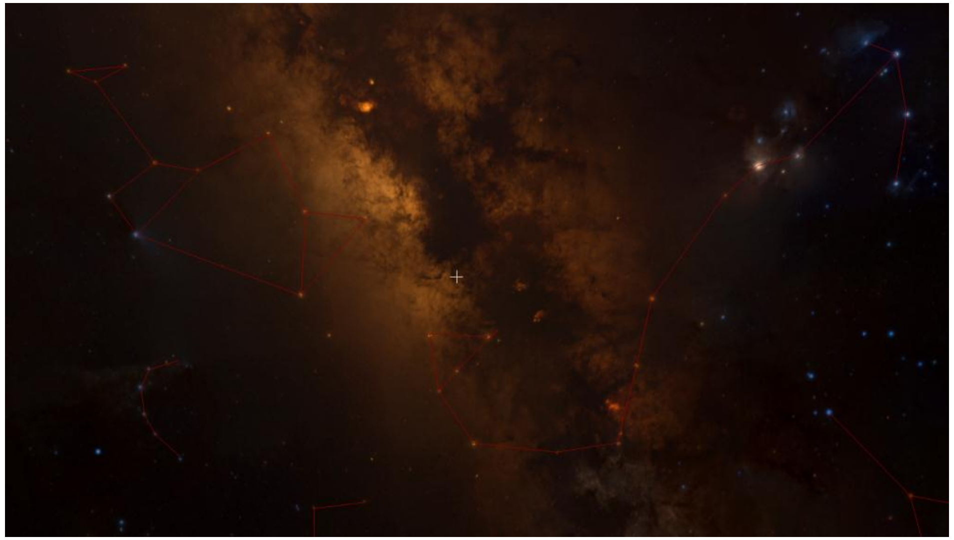
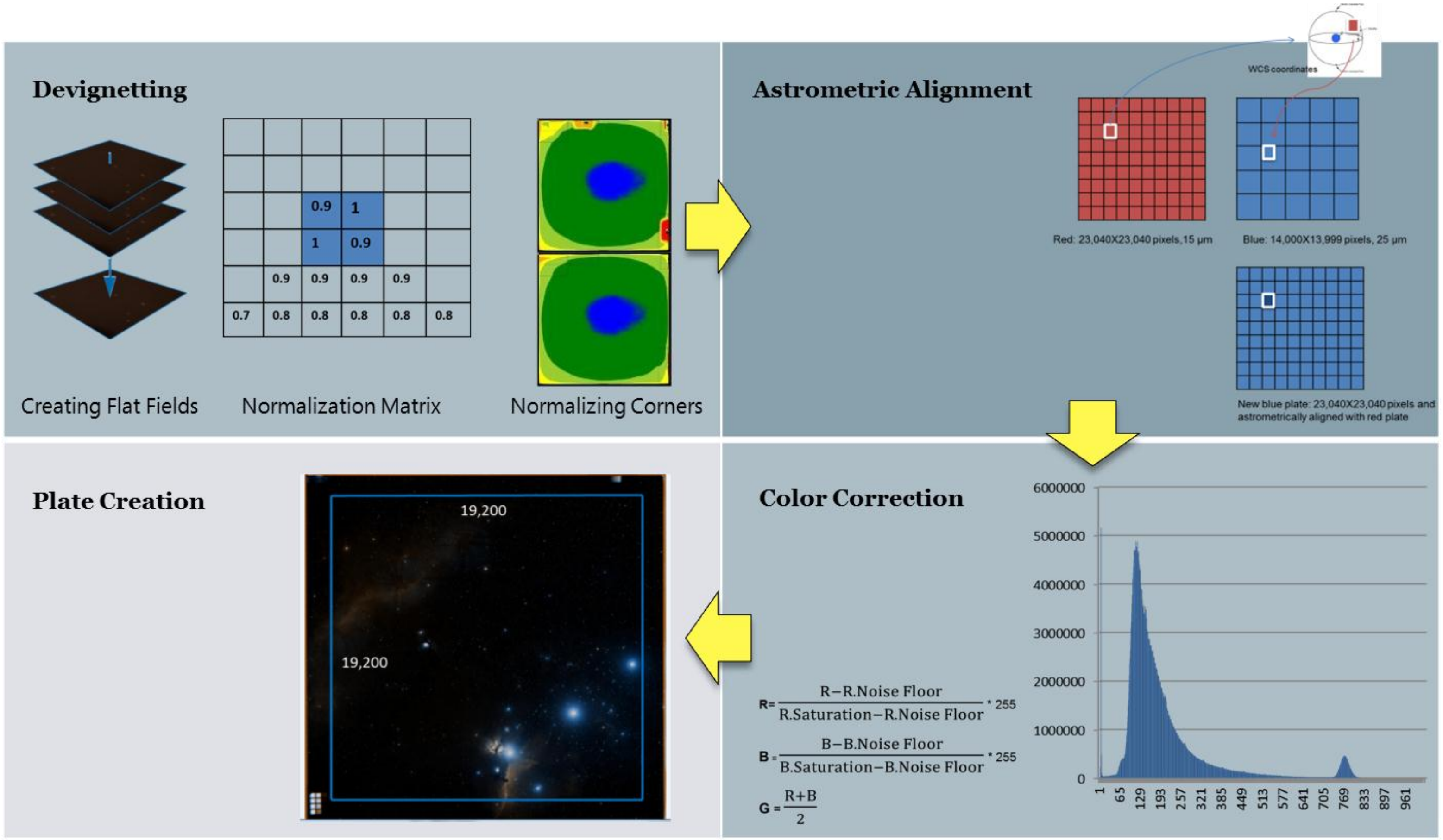


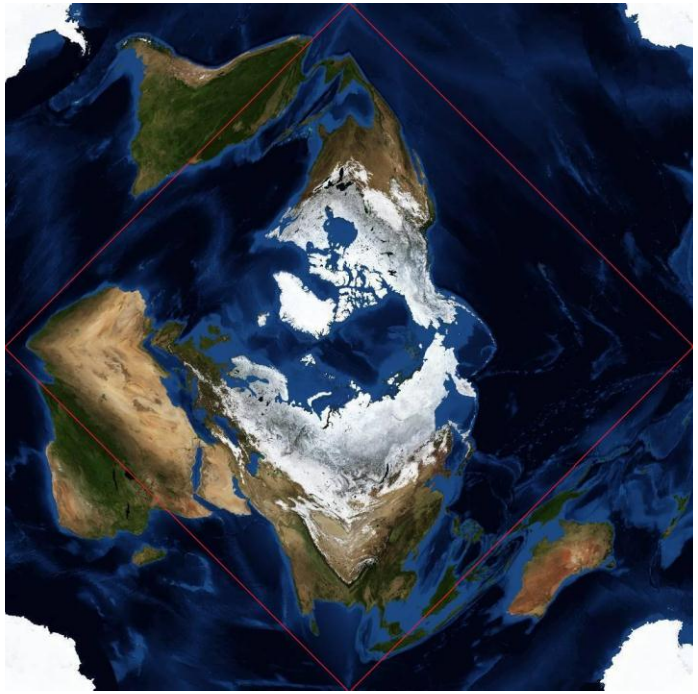
Image Correction and Plate Generation

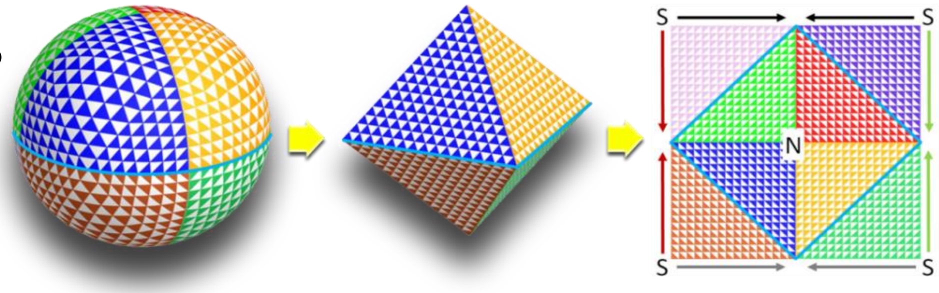
- Devignetting and artifact corrections
  - Correcting edge and corner darkening and varying levels in brightness, noise and saturation.
  - Creating a matrix of per pixel normalized correction factors and programmatically normalizing at selected regions.
- Astrometric Alignment
  - Generation of a new blue plate that has the same pixel granularity and location as the red plate
- Color Correction
  - Applying saturation and noise floor to red and blue channels; generating a green channel
- Plate Creation
  - Each color image is cropped to 19,200 x 19,200 pixels



TOAST Reprojection and Stitching

- Color images reprojected to TOAST (Tesllated Octahedral Adaptive Subdivision Transform)
  - 3 bits per pixel sky view
  - 2 bits per pixel seam location mask
  - 13 level tile pyramid
- Stitched together to set all gradients across image boundaries to zero
  - Executes in lockstep across all processes

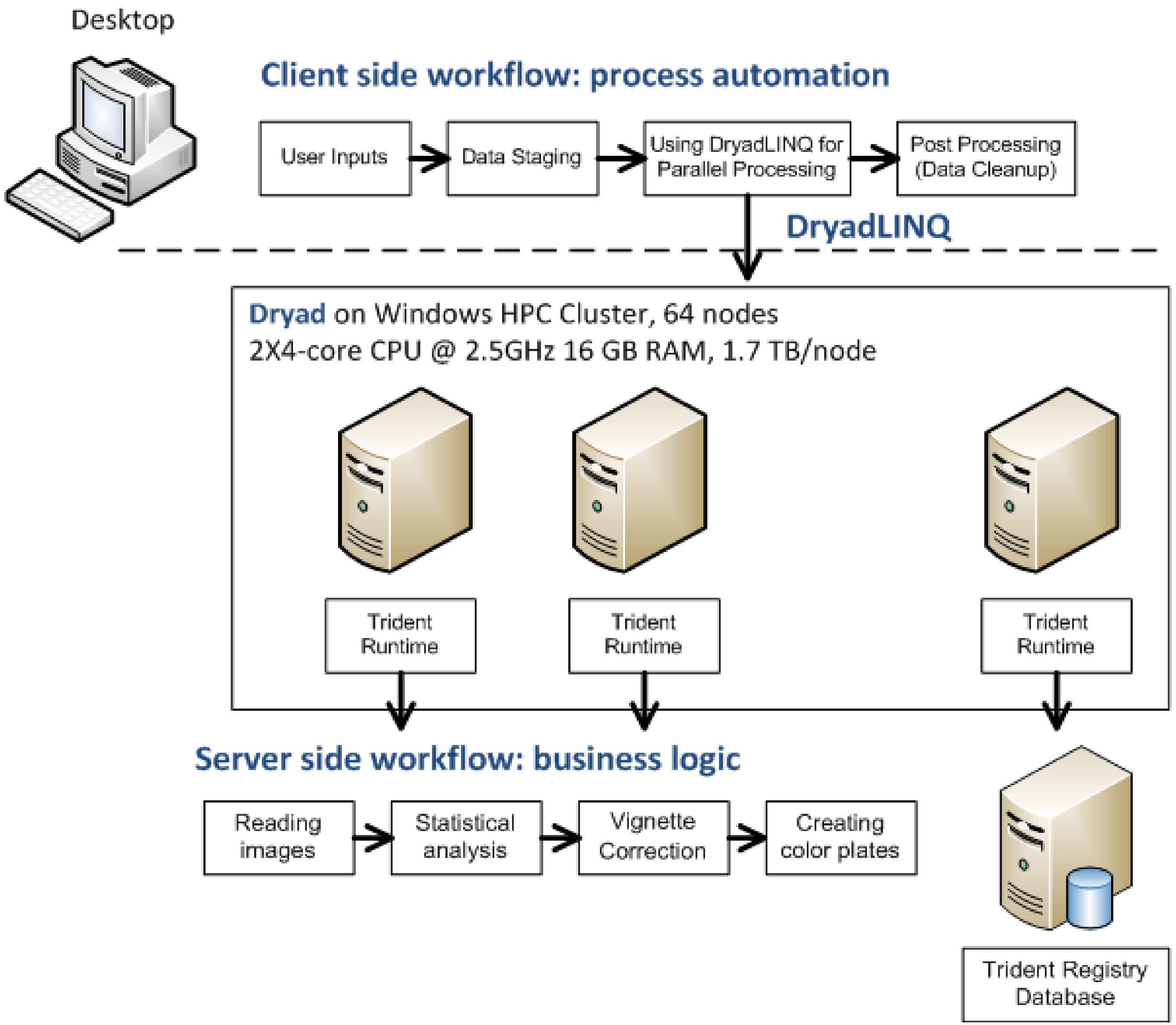




M. Kazhdan, D. Surendran, H. Hoppe. Distributed gradient-domain processing of planar and spherical images, *ACM Trans. on Graphics*, 29(2), 14, 2010.

Four Stage Image Processing Pipeline

- Trident Scientific Workflow Workbench manages the overall process from the desktop
- DryadLINQ and .NET parallel extensions manages the server execution
- Microsoft Windows HPC Server provides the basic scheduling and monitoring abstractions



Dryad

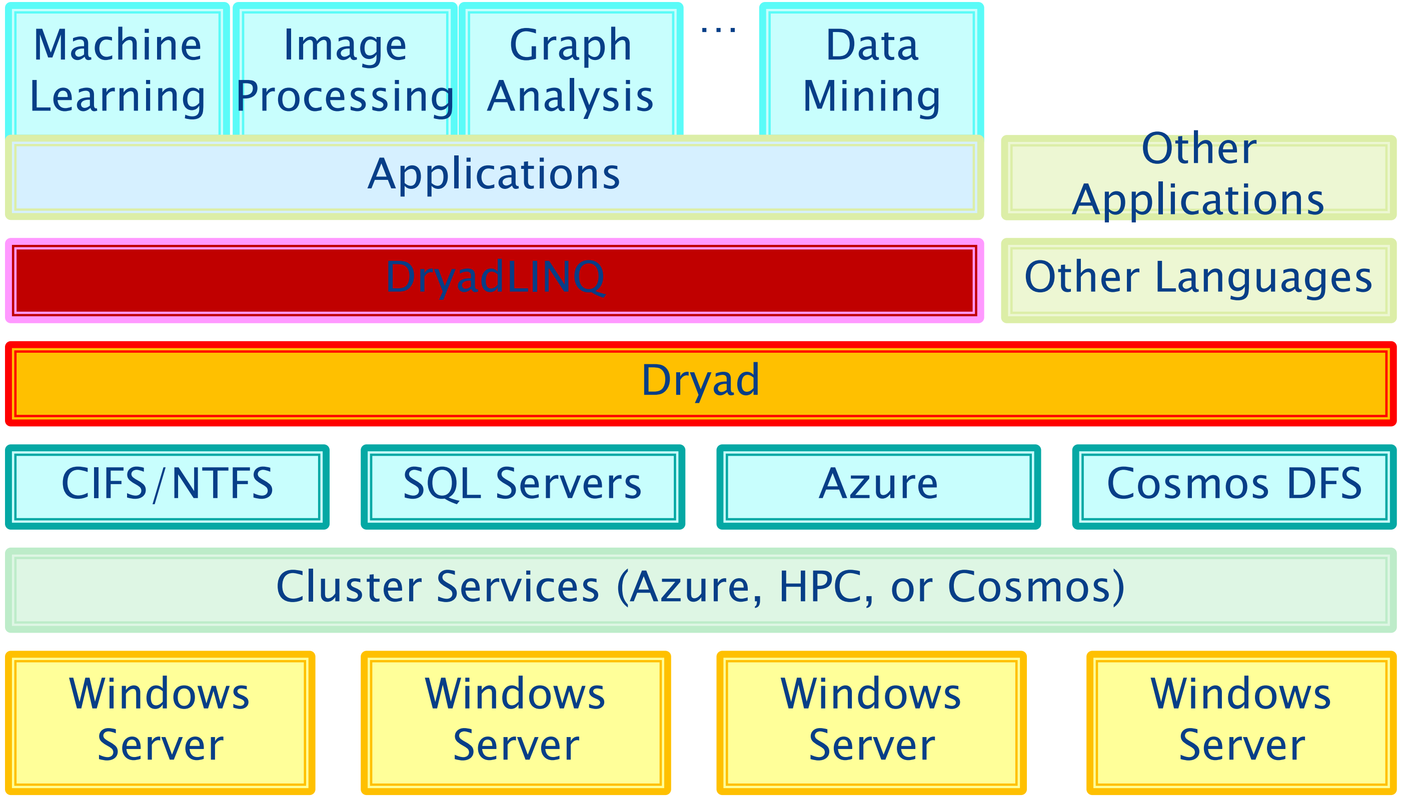
- Use a cluster as if it were a single computer
  - Sequential, single machine programming abstraction
  - Same program runs on single-core, multi-core, or cluster
- Continuously deployed since 2006
  - The execution engine for Bing
  - $>10^4$ machines with single clusters > 3000 machines
  - Sifting through datasets >10 PB daily
- Familiar programming languages and development environment
  - C#, VB, F#, IronPython…with .NET, Visual Studio or other IDE

http://connect.microsoft.com/dryad

http://research.microsoft.com/collaboration/tools/dryad.aspx

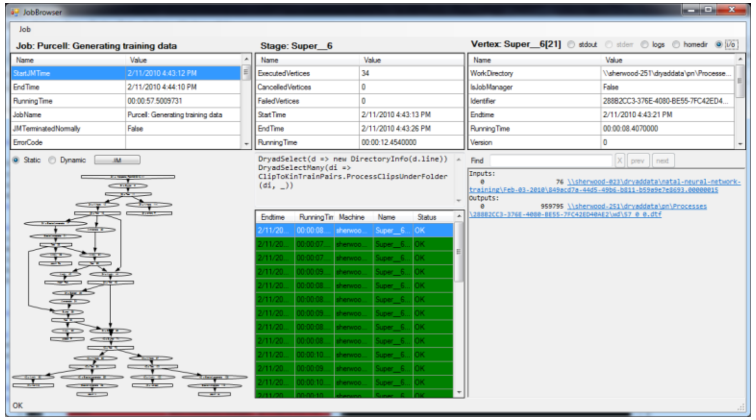
LINQ

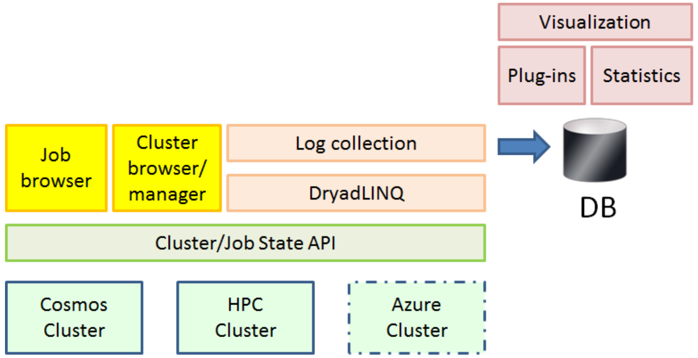
- Microsoft's Language INtegrated Query
- A set of operators to manipulate datasets in .NET
  - Support traditional relational operator such as select, Join, GroupBy, Aggregate, etc.
  - Integrated into .NET programming languages: programs can call operators and operators can invoke arbitrary .NET functions
- Data model
  - Data elements are strongly typed .NET objects
  - Much more expressive than SQL tables
- Extremely extensible
  - Add new custom operators
  - Add new execution providers

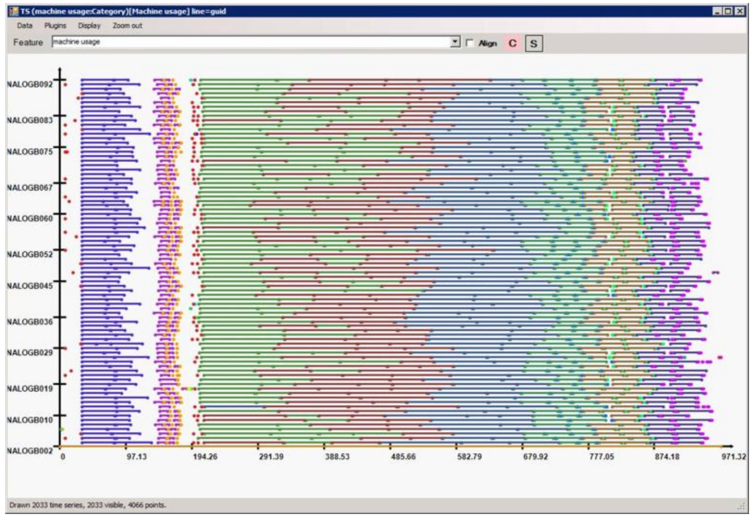


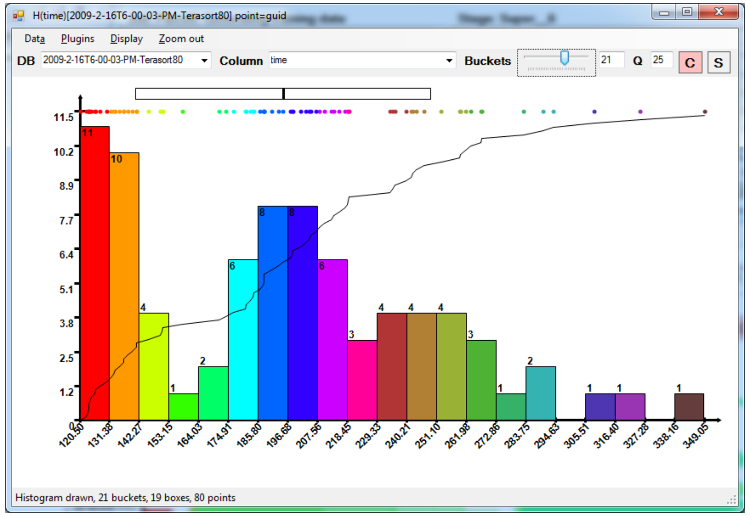
DryadLINQ

- Distributed execution plan generation
  - Static optimizations: pipelining, eager aggregation, etc.
  - Dynamic optimizations: data-dependent partitioning, dynamic aggregation, etc.
- Vertex runtime
  - Single machine (multi-core) implementation of LINQ
  - Vertex code that runs on vertices
  - Data serialization code
  - Callback code for runtime dynamic optimizations
  - Automatically distributed to cluster machines





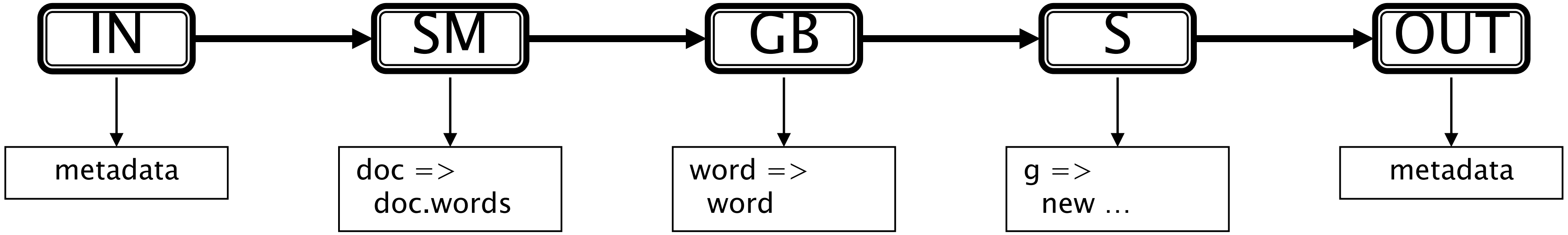




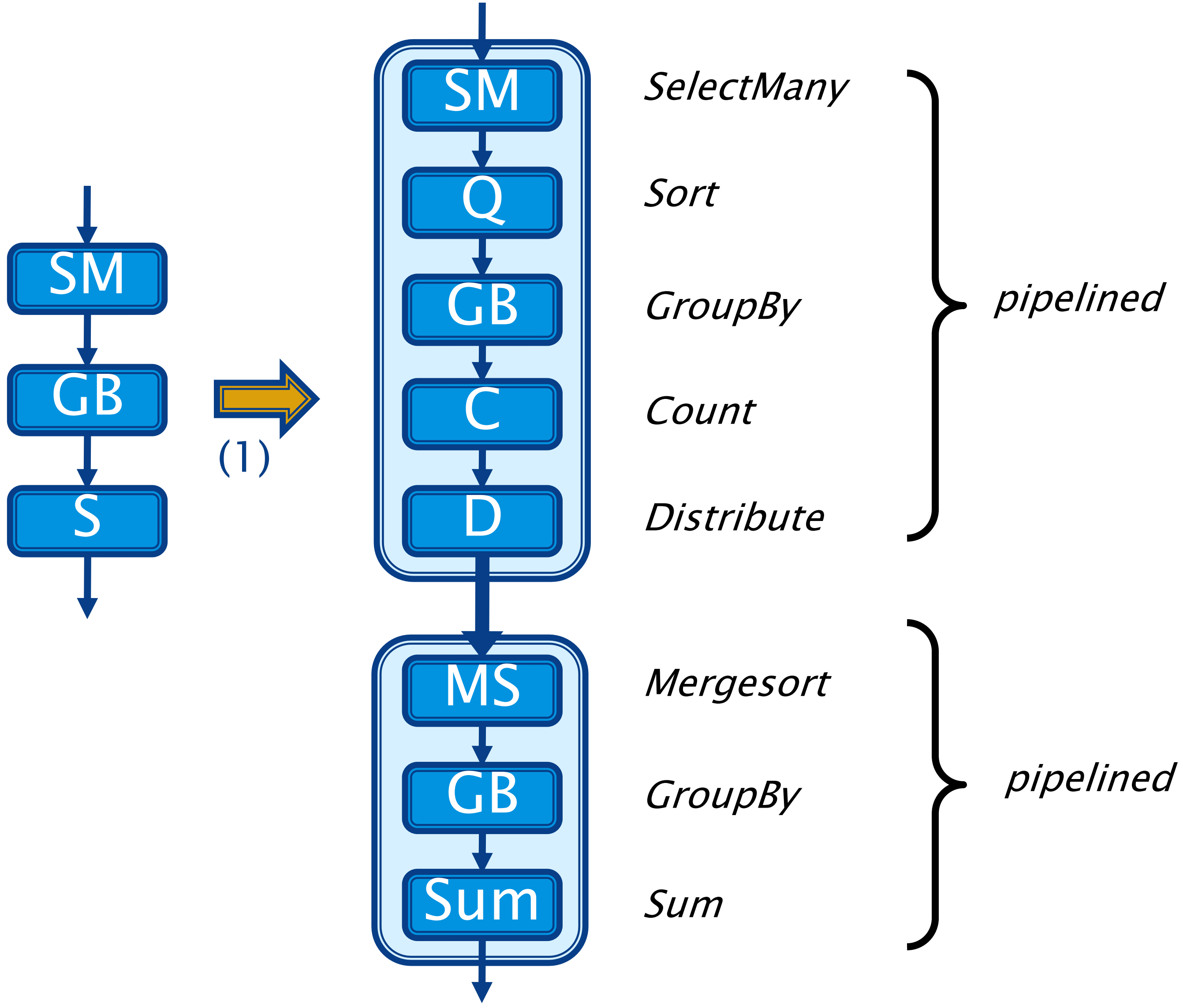
Example: Word Count

Count word frequency in a set of documents:

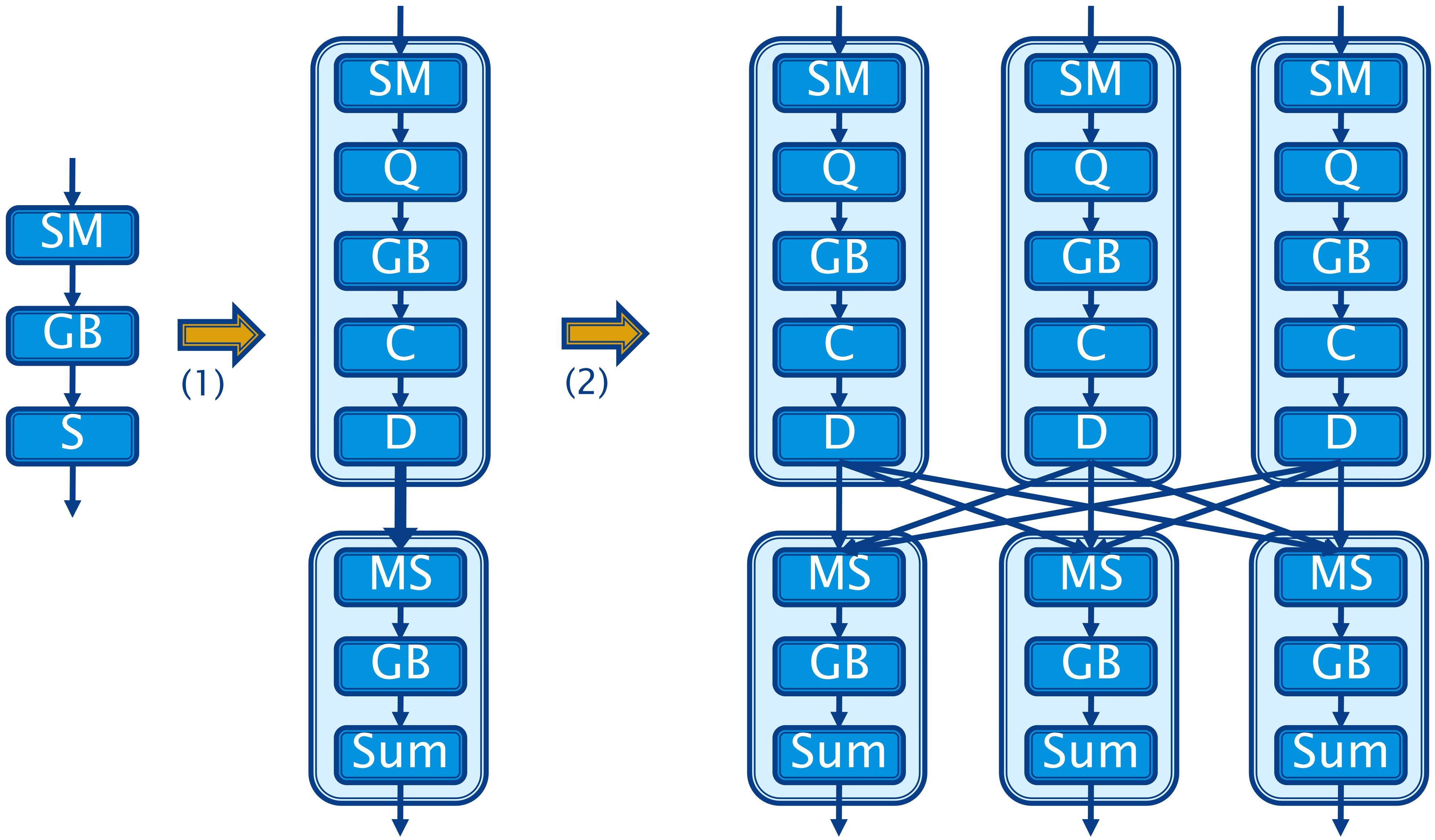
```
var docs = new PartitionedTable<Doc>("dfs://yuan/docs");
var words = docs.SelectMany(doc => doc.words);
var groups = words.GroupBy(word => word);
var counts = groups.Select(g => new WordCount(g.Key, g.Count()));
counts.ToTable("dfs://yuan/counts.txt");
```



Execution Plan for Word Count



Execution Plan for Word Count



Vignetting for Image Normalization

- Three sequential algorithms:
  - `ImageToRows`: loads, normalizes and shreds an image into a row of pixels
  - `ReduceStackRows`: per pixel averaging across plates
  - `SaveFlatField`: persist results
- PartitionedTable Creation to hold source imagery

```
Parallel.For(0, ny, y => {
     for (int x=0; x<nx; x++)
          // do work at (x,y)
});
```

- DryadLINQ to distribute computation

```
var rows = images.SelectMany(image =>
         ImageToRows(image, options));
var stackedRows =
         pixelRows.GroupBy(row =>
         row.Position);
var finalRows = stackedRows.Select(x =>
     ReduceStackedRows(x));
var flatfield = finalRows.Apply(x =>
     SaveFlatField(x, options));
```

TeraPixel By the Numbers

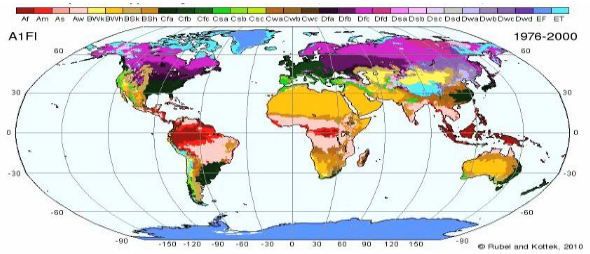
- Input: 1791 pairs of red blue images:
  - 417 GB
- Output: 1025 full color 24-bit TOAST pyramid files:
  - 800 GB.
- Cluster: 64 compute nodes 8 core Intel Xeon, 16 GB RAM, 1.7 TB storage, 1 Gbps link
- Generation of RGP plates
  - 5 hours processing
- Image stitching into a spherical image
  - 3 hours processing
- Image optimization to remove seams
  - 4.5 hours processing
- Results staging off cluster
  - 2.5 hours
- The resulting image:
  - 24 bit RGB terapixel image of the night sky.
  - 500,000 HDTVs to view image at full resolution
  - A football field sized paper to print the image

MODISAzure

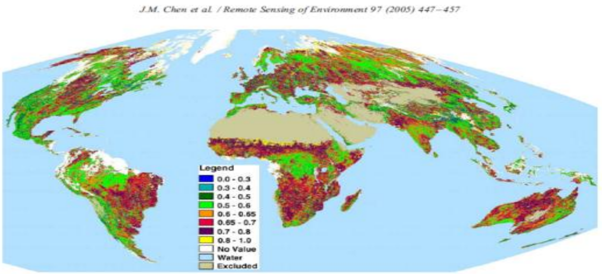
Behind every cloud is another cloud.

Judy Garland

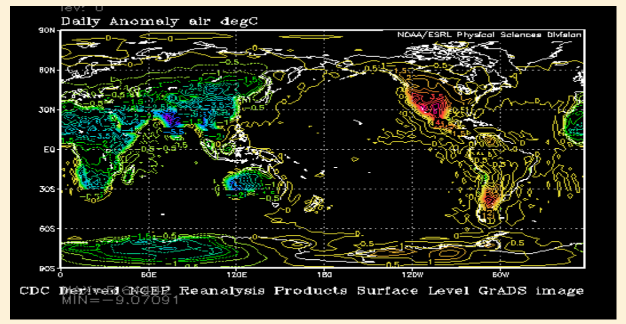
Synthesizing Imagery, Sensors and Field Data



Climate classification
~1MB (1file)

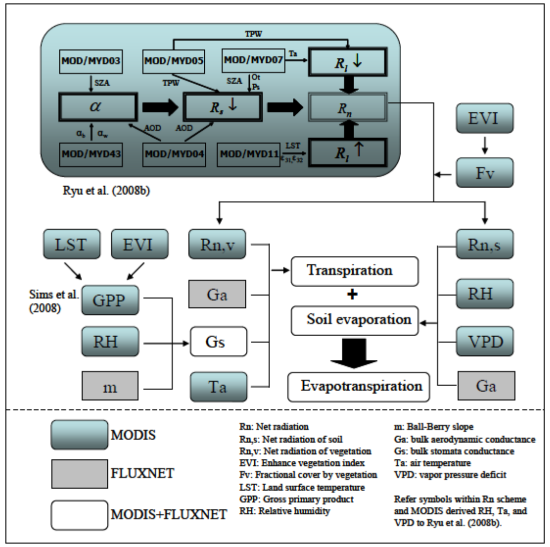


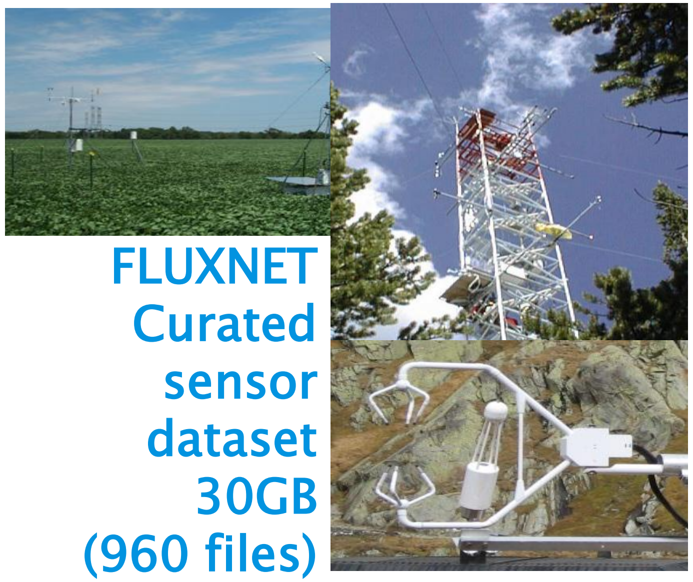
Vegetative clumping
~5MB (1file)



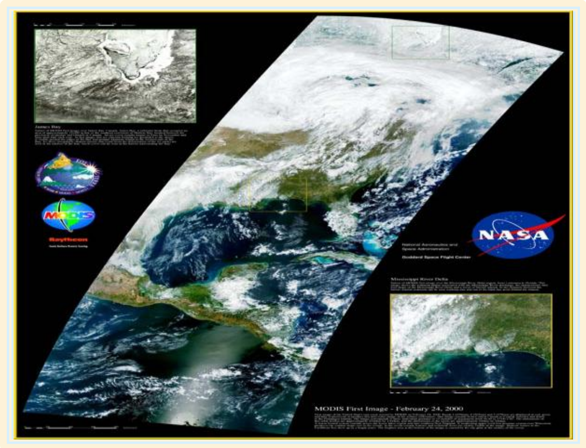
NCEP/NCAR ~100MB
(4K files)

Not just a simple matrix computation due to dry region leaf/air temperatures differences, snow cover, leaf area fill, temporal upscaling, gap fill, biome conductance lookup, C3/C4 plants, etc etc



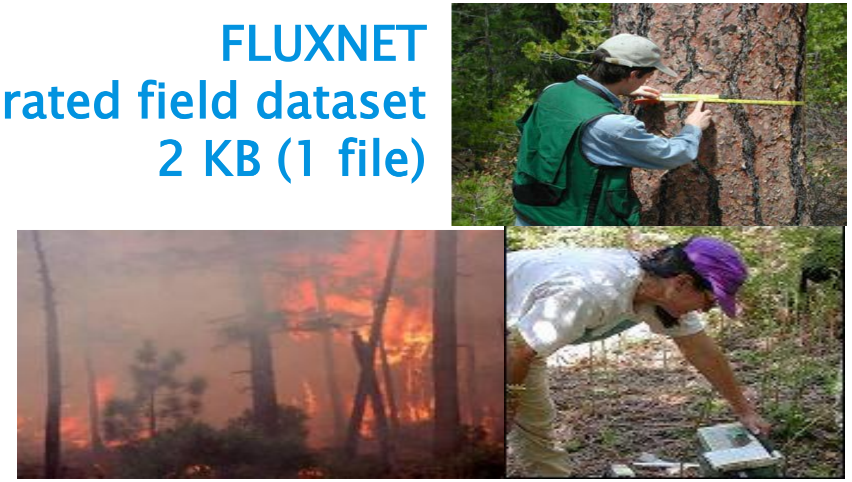


FLUXNET
Curated
sensor
dataset
30GB
(960 files)



NASA MODIS imagery archives
5 TB (600K files) for 10 US years

FLUXNET
curated field dataset
2 KB (1 file)

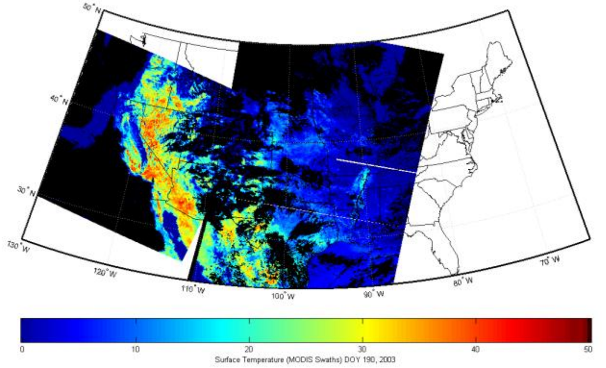


The Tedium Factor: Do Scientists have to become Computer Scientists?

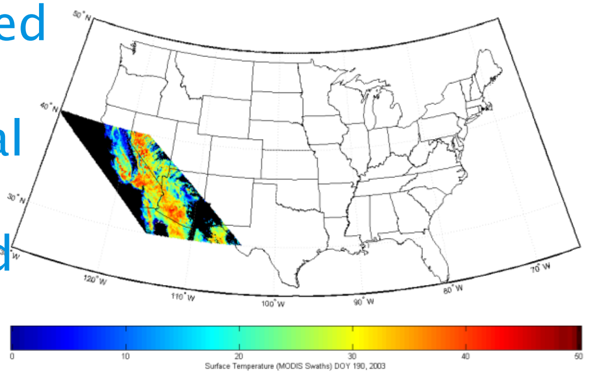
- Downloading
  - Example: identifying and downloading the swath precursors necessary to reproject a given sinusoidal cell
- Reprojection.
  - Example: latitude-longitude swaths to sinusoidal cells.
- Spatial resampling .
  - Example: converting from 1 KM to 5 KB pixels.
- Temporal resampling
  - Example: converting from daily observation to 8 day averages.
- Gap filling
  - Example: assigning values to pixels without data due to cloud or satellite outages.
- Masking
  - Examples: eliminating pixels over the ocean when computing land product or outside a spatial feature such as a watershed.

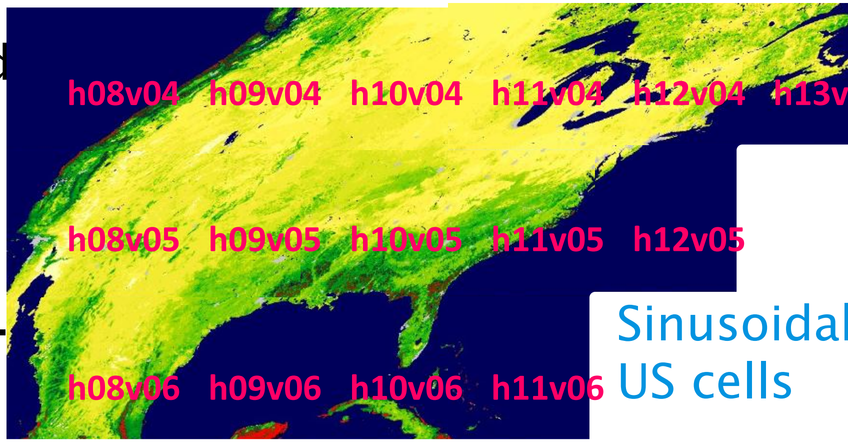
Grunge means you're doing science

Source Data (Swath format)



Reprojected Data (Sinusoidal format – equal land area pixel)

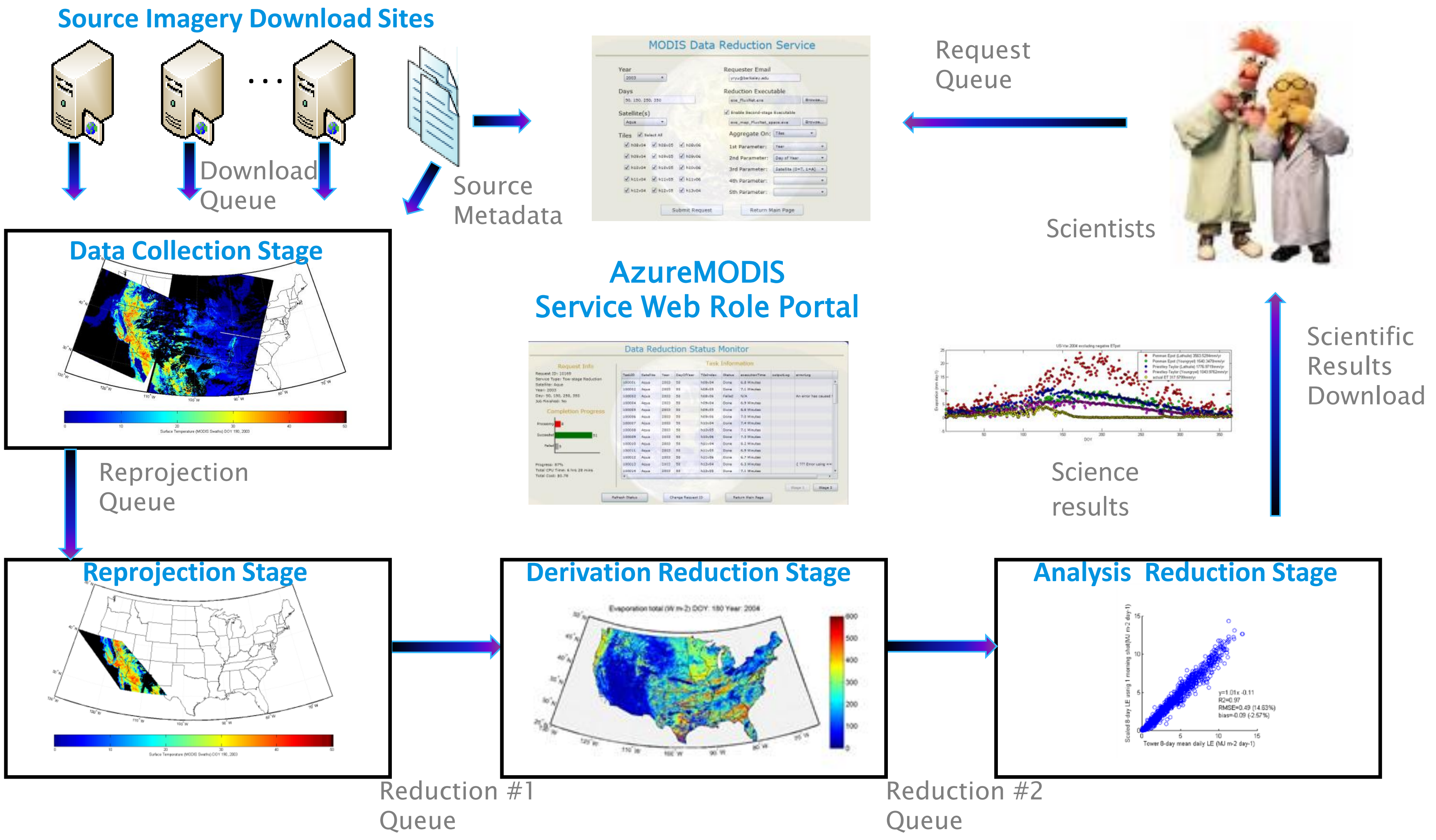




Sinusoidal US cells

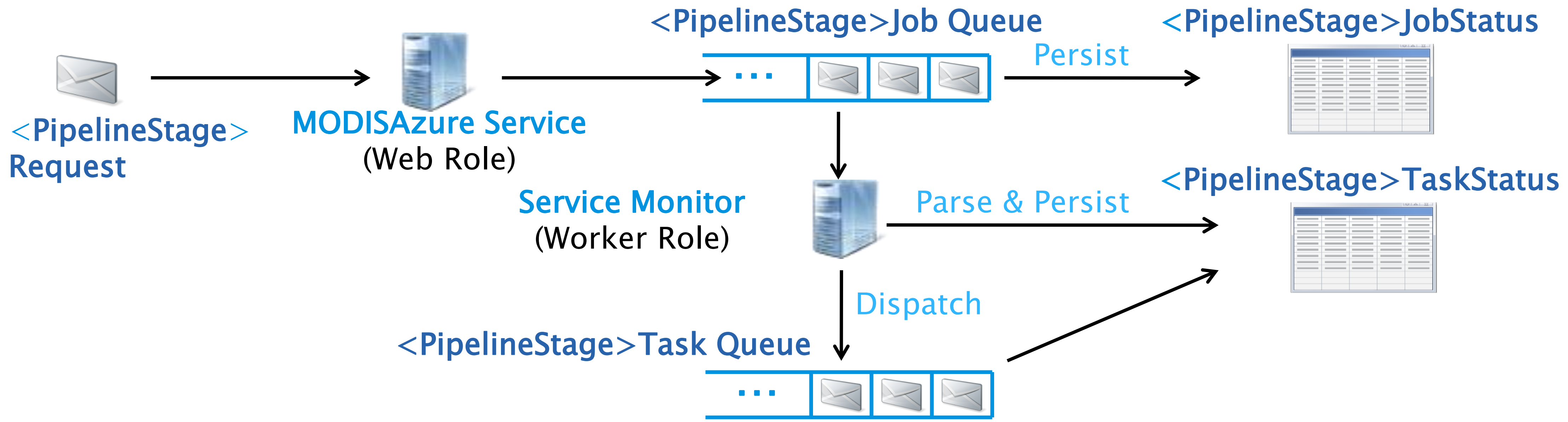
Four Stage Image Processing Pipeline

- Data collection stage
  - Downloads requested input tiles from NASA ftp sites
  - Includes geospatial lookup for non-sinusoidal tiles that will contribute to a reprojected sinusoidal tile
- Reprojection stage
  - Converts source tile(s) to intermediate result sinusoidal tiles
  - Simple nearest neighbor or spline algorithms
- Derivation reduction stage
  - First stage visible to scientist
  - Computes ET in our initial use
- Analysis reduction stage
  - Optional second stage visible to scientist
  - Enables production of science analysis artifacts such as maps, tables, virtual sensors



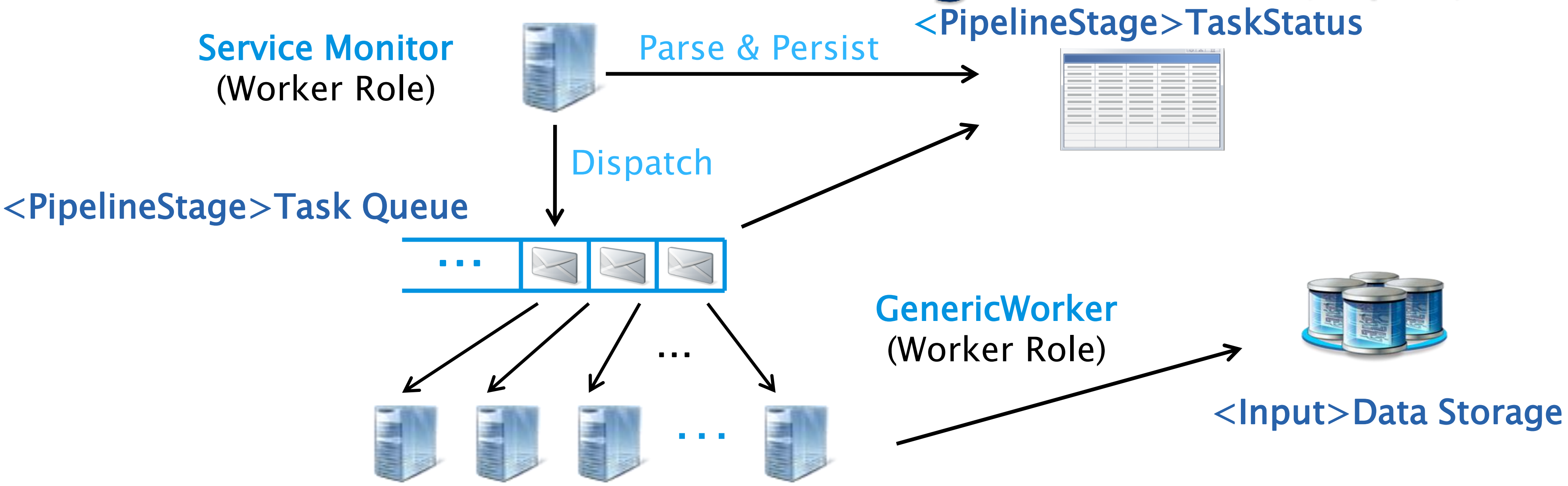
http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx

MODISAzure: Architectural Big Picture (1/2)



- **ModisAzure Service** is the *Web Role* front door
  - Receives all user requests
  - Queues request to appropriate Download, Reprojection, or Reduction Job Queue
- **Service Monitor** is a dedicated *Worker Role*
  - Parses all job requests into tasks – recoverable units of work
  - Execution status of all jobs and tasks persisted in Tables

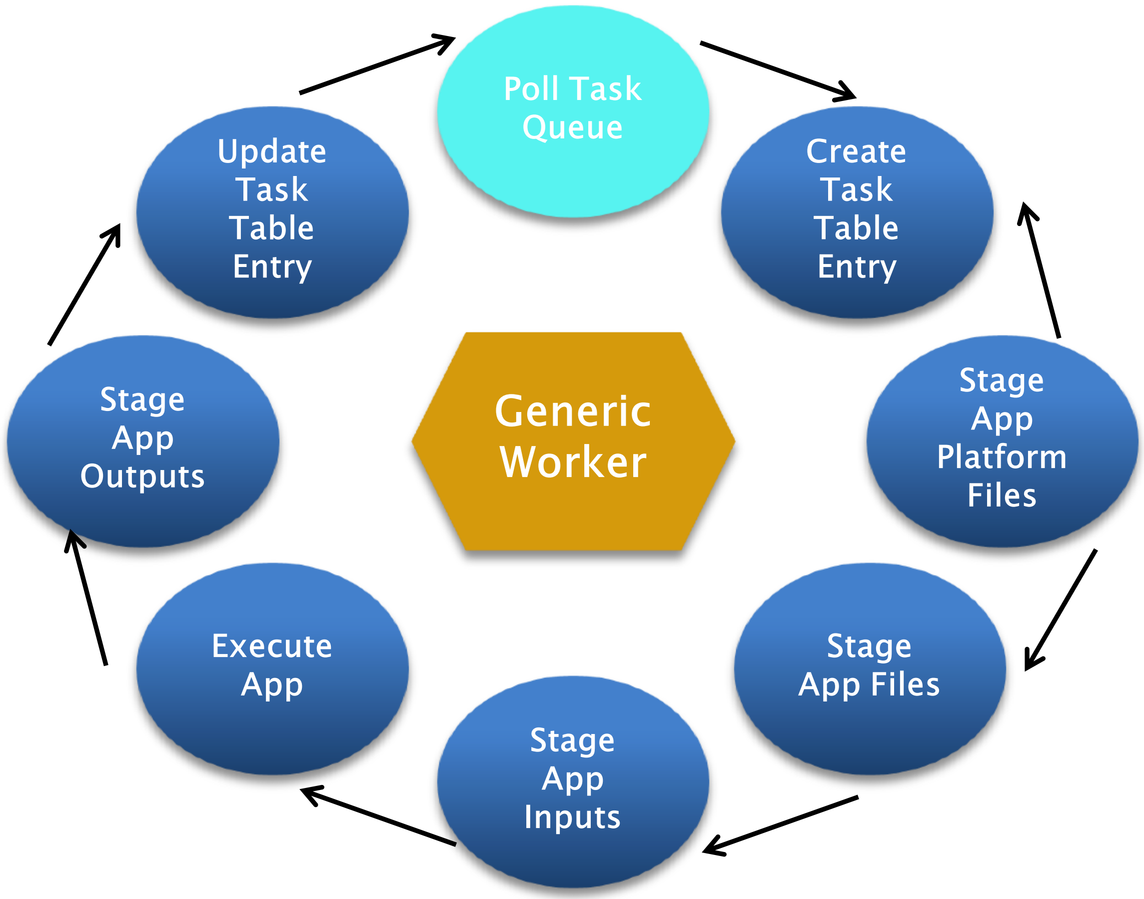
MODISAzure: Architectural Big Picture (2/2)



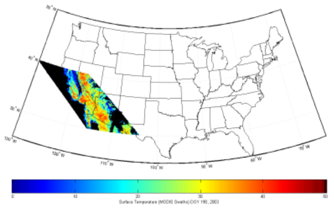
- All work actually done by a GenericWorker *Worker Role*
  - Dequeues tasks created by the Service Monitor
  - Science executable is sandboxed on an Azure Worker instance thereby enabling simple desktop development and debug
  - Marshalls all storage from/to Azure blob storage to/from local Azure Worker instance files
  - Retries failed tasks 3 times
  - Maintains all task status

Inside A Generic Worker

- Manages application sandbox
  - Ensures all application binaries such as the MatLab runtime are installed for “known” application types
  - Stages all input blobs from Azure storage to local files
  - Passes any marshalled inputs to uploaded application binary
  - Stages all output blobs to Azure storage from local files
  - Preserves any marshalled outputs to the appropriate Task table
- Manages all task status
  - Dequeues tasks created by the Service Monitor
  - Retries failed tasks 3 times
  - Maintains all task status
- Simplifies desktop development and cloud deployment



Determining What to Download



- Each product is either swath or sinusoidal projection
  - Sinusoidal are ready to use
  - Groups of swath products must be reprojected to create a sinusoidal tile
- NASA publishes a geometadata information for the two Terra and Aqua satellites
- For each 5 minute swath data file (or granule) on the ftp site there is a corresponding geometa file containing: DayNightFlag indicating day, night or both; corner point latitude/longitude; bounding coordinates
- We ingested all files (288 per day * 10 years * 2 satellites) into a SQL database then paged the information into our Azure ScanTimeList and GeoMeta Tables
- The dayScanTimeList in the ScanTimeList table identifies all swath source file precursors for a given sinusoidal tile and drives the download and reprojection

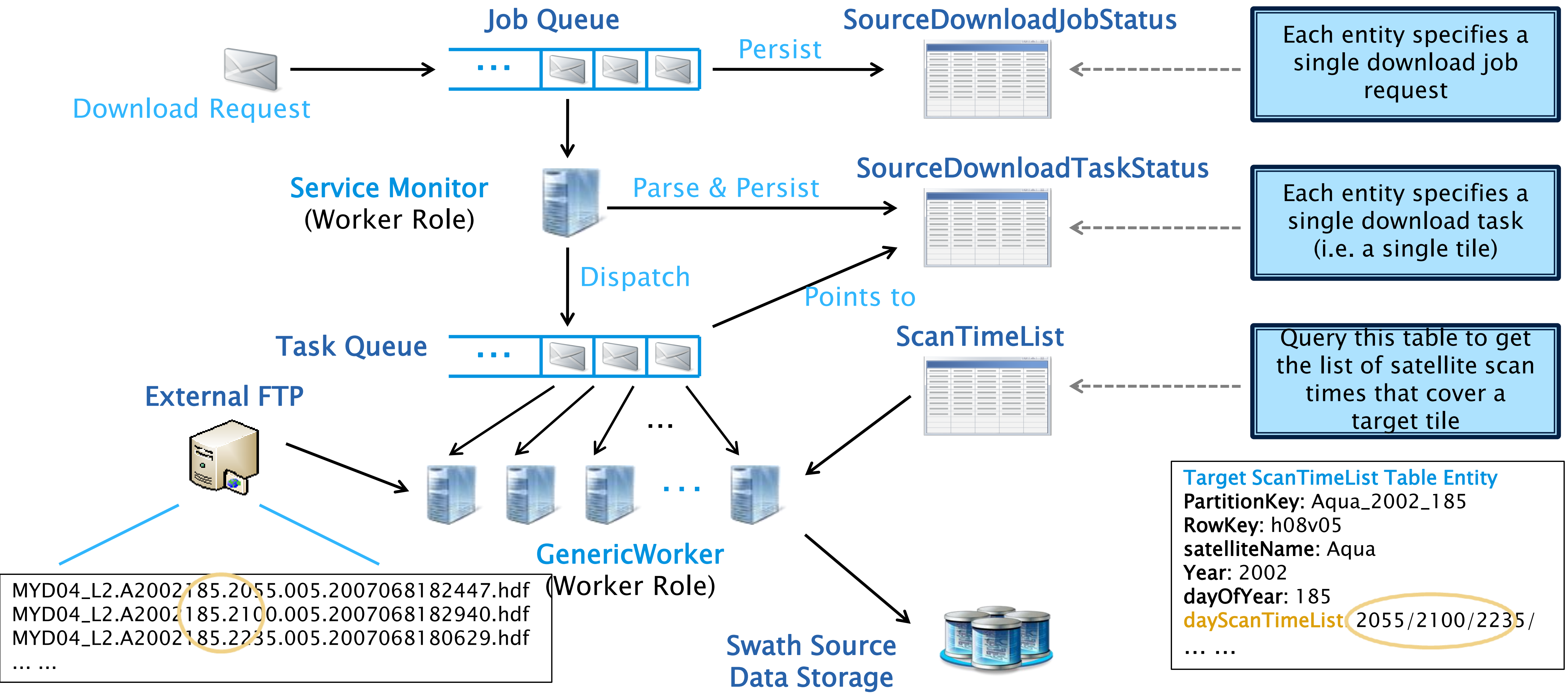
M*D04	Aerosol
M*D05	Precipitable water
M*D06	Cloud
M*D07	Temperature, ozone
MCD43B*	Albedo
M*D11	Surface temperature
M*D15	LAI
MOD13A2	Vegetation Index
MCD12Q1	Land Cover
MOD44B	Veg. Contig. Fields

#Attributes PartitionK	RowKey	Timestamp	betweenScanTimeList	dayOfYear	dayScanTimeList	hIndex	nightScanTimeList	satellite	vIndex	year
Terra_2003_160	h00v07	2/10/2010 7:33		160	2220/2355/	0	1005/1010/1145/	Terra	0	2003

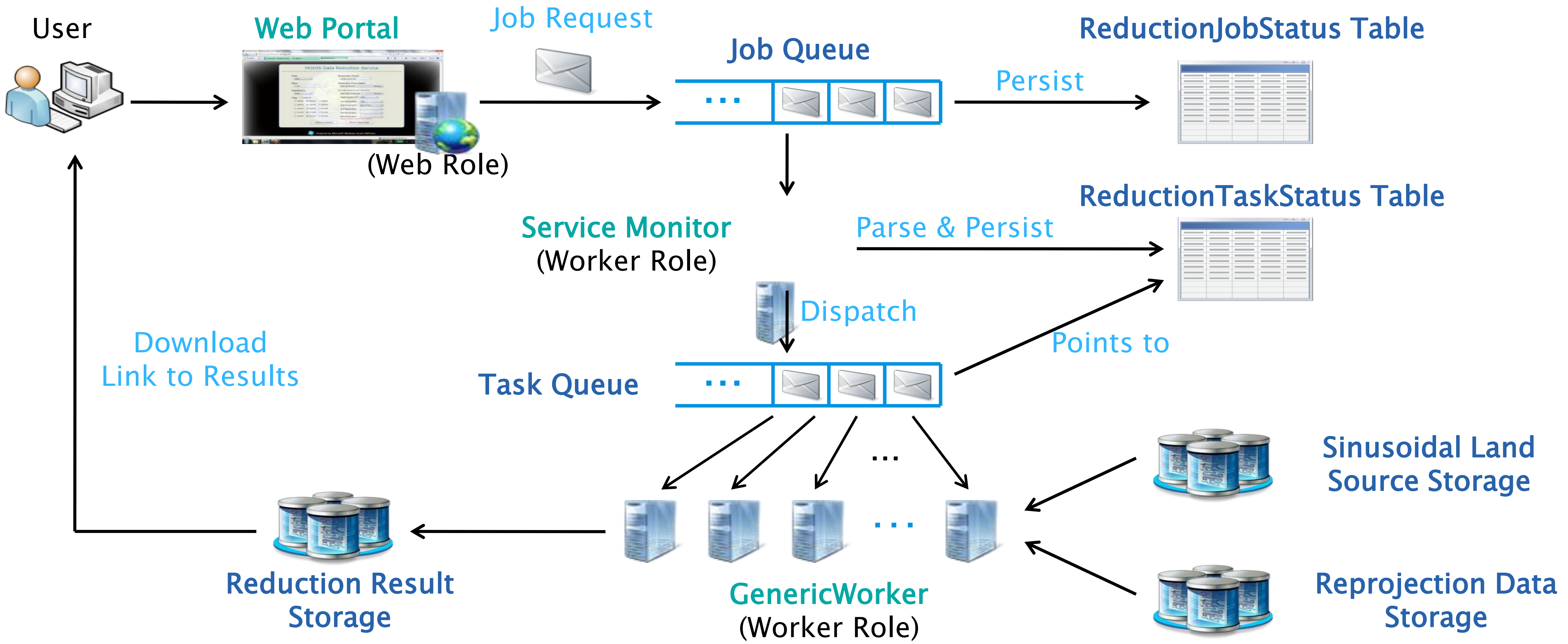
ftp://ladsweb.nascom.nasa.gov/geoMeta/README

Source Data Download Service

Example: Download the required source files for the target sinusoidal tile: **MYD04_L2, Year 2002, Day 185, h08v05**

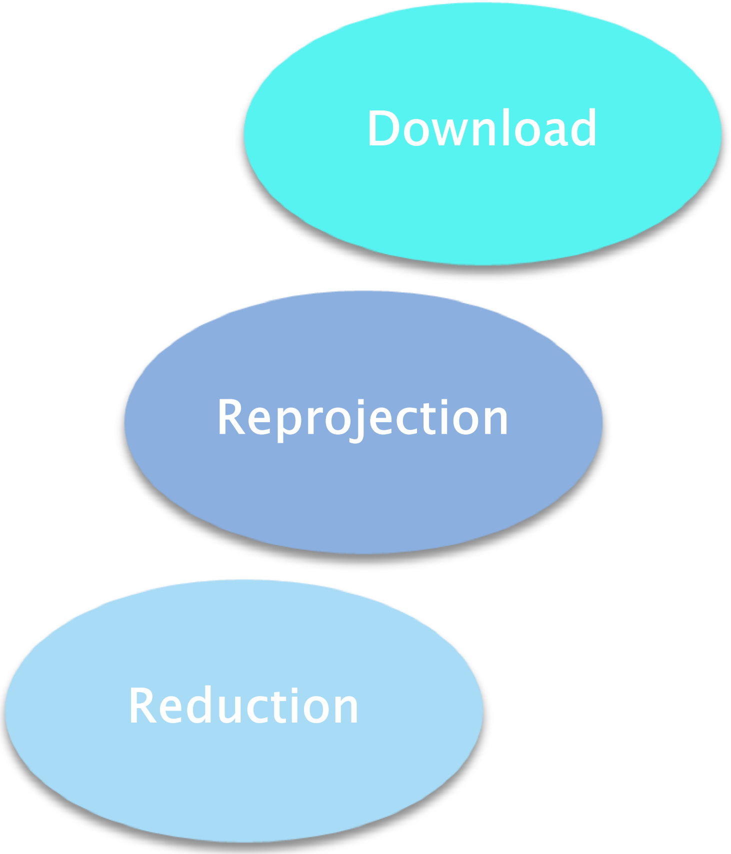


Reduction Service (Only One Stage Shown)



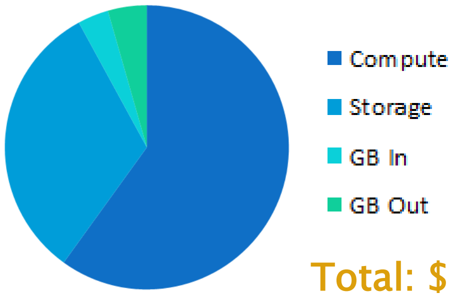
Pipeline Stage Interactions

- The Web Portal Role, Service Monitor Role and 5 Generic Worker Roles are deployed at most times
  - 5 Generic Workers are sufficient for reduction algorithm testing and development ($20/day)
  - Early results returned to scientist while deploying up to 93 additional Generic Workers; such a deployment typically takes 45 minutes
  - Deployment taken down when long periods of idle time are known
  - Heuristic for scaling number of Generic Workers up and down
- Download stage runs in the deep background in all deployed generic worker roles
  - IO, not CPU bound so no competition
- Reduction tasks that have available inputs run preferentially to Reprojection tasks
  - Expedites interactive science result generation
  - If no available inputs and a backlog of reprojection tasks, number of Generic Workers scale up naturally until backlog addressed and reduction can continue
  - Second stage reduction runs only after all first stage reductions have completed
- Reduction results can be downloaded following emailed link to zip file

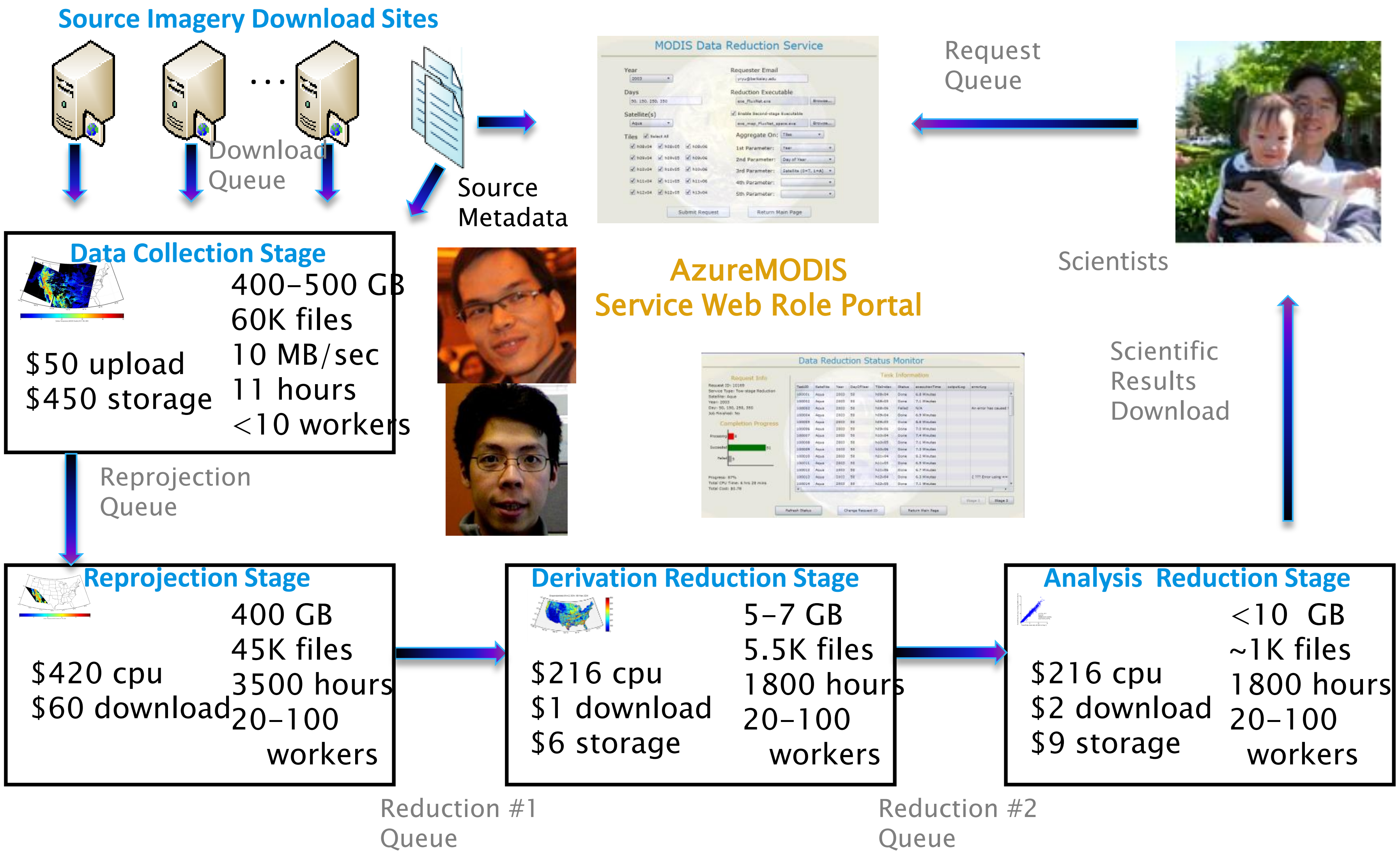


Costs for 1 US Year ET Computation

- Computational costs driven by data scale and need to run reduction multiple times
- Storage costs driven by data scale and 6 month project duration
- Small with respect to the people costs even at graduate student rates !

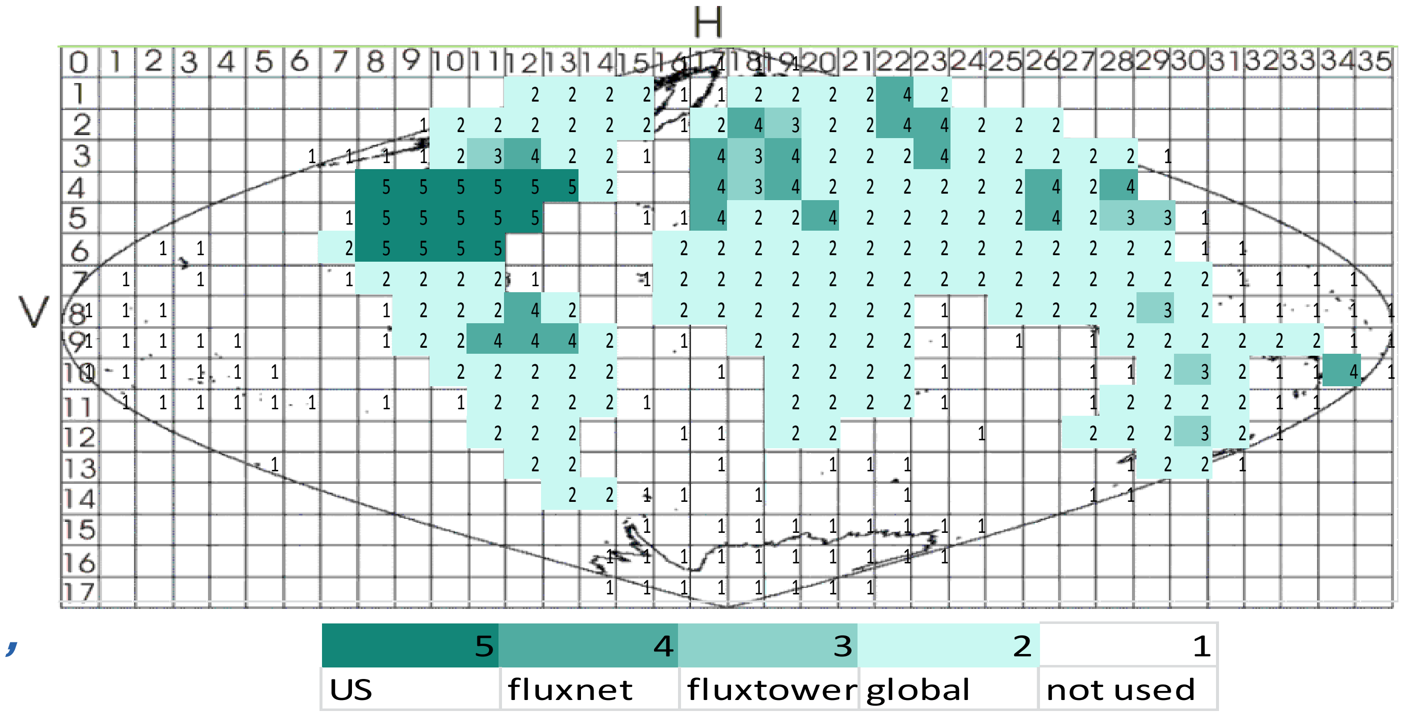


Total: $1420



MODISAzure Global Computation By the Numbers

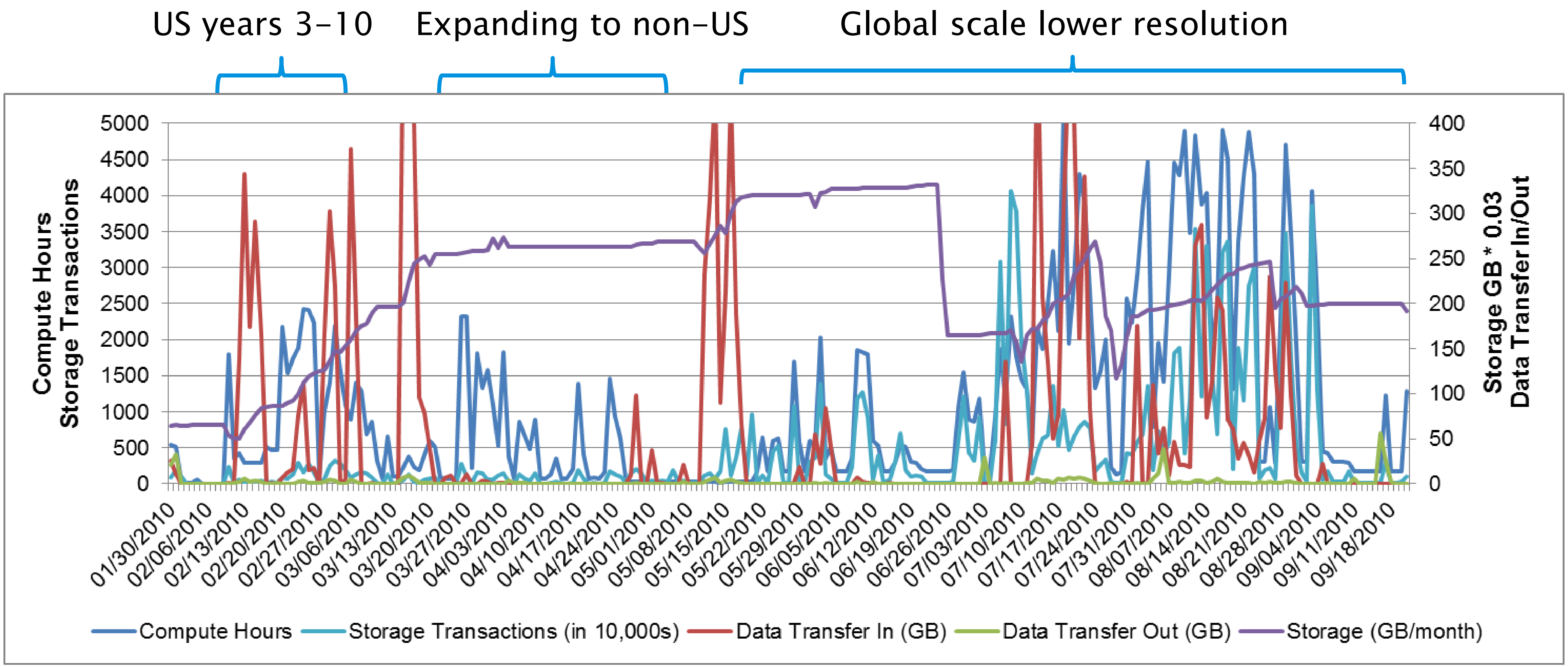
- 194 sinusoidal cells, each covers 1.2x1.2 KM or 11M 5 KM pixels)
- 1.06 M reprojected tiles and 40.5K source sinusoidal tiles
- 8 TB (>10 M files) downloaded from NASA ftp
- Not all files are downloaded or reprojected at first (3 rapid retries) attempt or actually available due to satellite outage, polar winter, missing tiles, etc etc.
- 55 NASA download days
- 150K reprojection compute hours
- 940 TB moved across Azure fabric
- 10 download result days (est) via IN2 bridge



15 seconds on the Cray Jaguar (1.75 PFLOPs) , but only if we could get the PB in

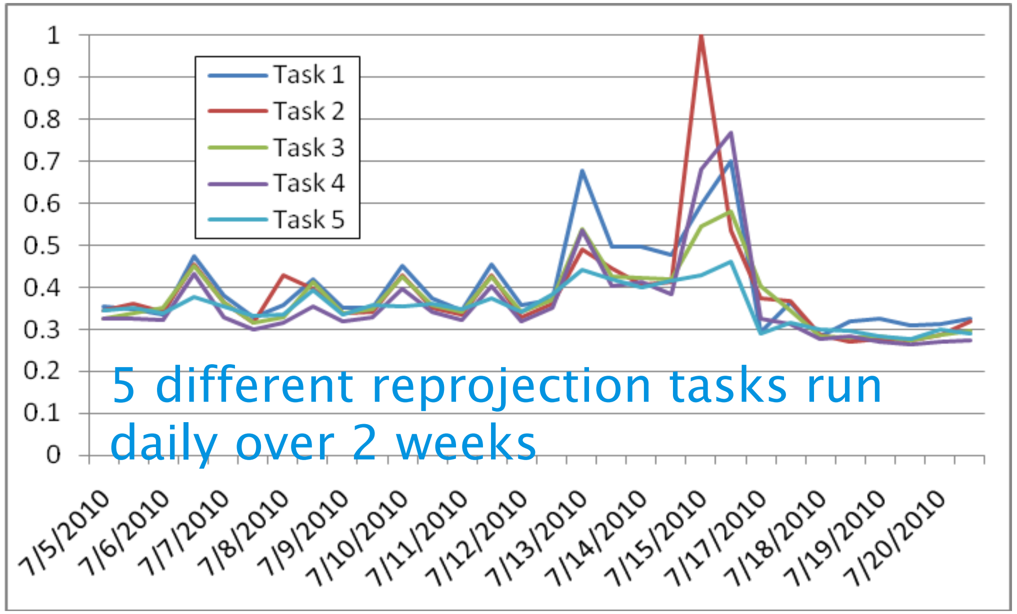
Agility

- The computation changed over time while Azure just scaled

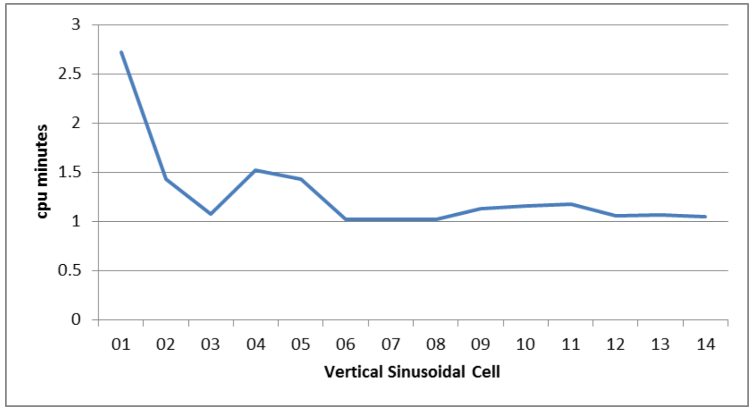


Predictability

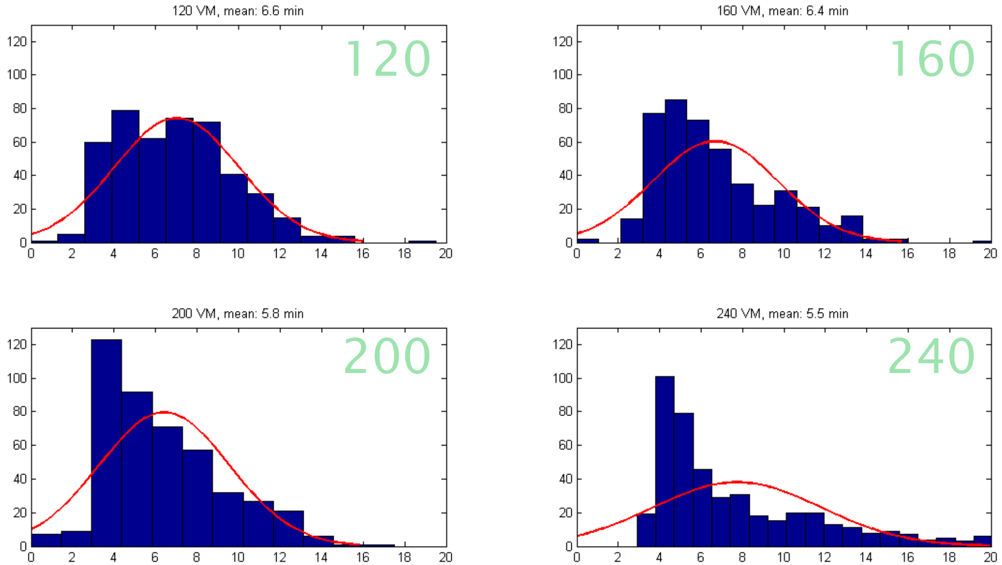
- Performance varies over time: rerunning the same task gives different timings on different days
- Performance varies over space: satellites are over the poles mo~~re~~ often



5 different reprojection tasks run daily over 2 weeks



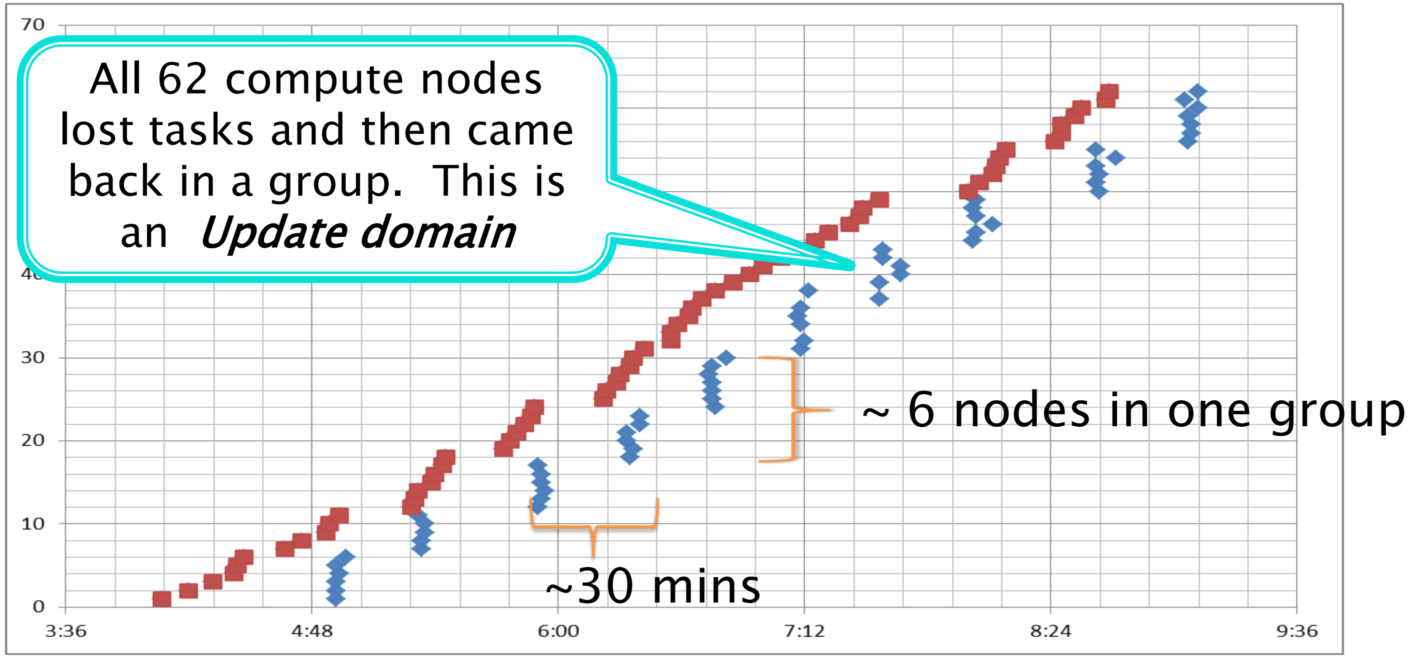
Average reprojection time (after algorithm improvements!) as a function of longitude



The same reduction task run on different numbers of VMs

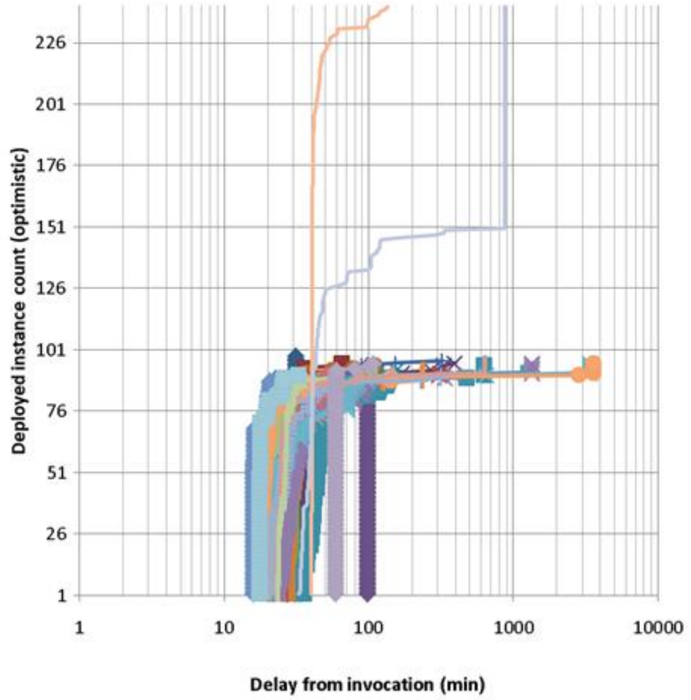
Reliability

- Even with 99.999% reliability, bad things happen
  - 1–2 % of MODISAzure tasks fail but succeed on retry

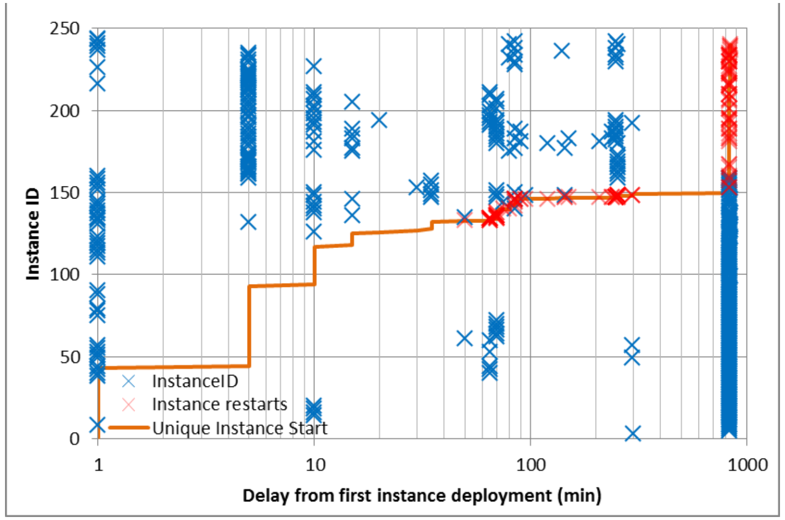


From AzureBlast
http://research.microsoft.com/en-us/people/barga/faculty_summit_2010.pdf

Observed VM starts for 76-100 VMs



Worst case attempt to start 250 VMs

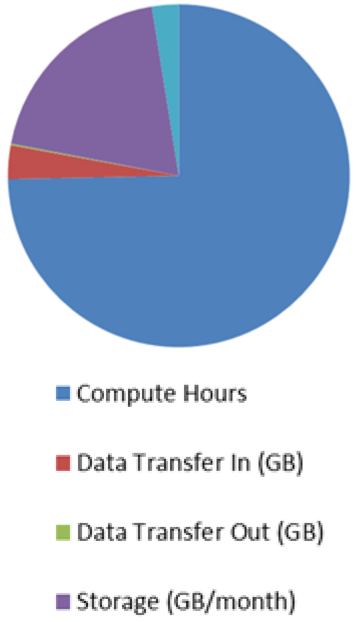


Fiscal Responsibility

- Billing is daunting
  - Neither we nor our academic collaborators are used to seeing bills
  - How *should* we think about them?
  - No billing cap means constant monitoring
- Billing is confusing
  - Instances are billed when deployed even if actually idle so comparing our usage log to the bill is at best approximate
  - Daily storage costs are amortized over the billing cycle so you must guestimate end cost
  - While you can ask for a refund, that takes a verified support call outage and time.
  - Online bill is autogenerated so must be accessed manually (no email)

Event Date	Name	Type	Region	Resource	Consumed	Sub Region	Service	Component	Service Info 1	Service Info 2	Additional Info
4/15/2010	Windows Azure Compute			Compute Hours	75		Compute	MODIS Data Services(Modis Data Service)			ComputeSmall
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB)	0.001257	South Central US	Storage	MODIS Source Data Products			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer In (GB) – Off Peak	0.000001	South Central US	Storage	MODIS Source Data Products			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB) – Off Peak	0.000018	South Central US	Storage	MODIS Source Data Products			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB) – Off Peak	0.00044	South Central US	Storage	Reduction Results			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB) – Off Peak	0.00002	South Central US	Storage	Reprojection Results			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer In (GB)	0.000032	South Central US	Compute	MODIS Data Services(Modis Data Service)			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB)	0.000033	South Central US	Compute	MODIS Data Services(Modis Data Service)			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer In (GB) – Off Peak	0.000005	South Central US	Compute	MODIS Data Services(Modis Data Service)			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB) – Off Peak	0.000003	South Central US	Compute	MODIS Data Services(Modis Data Service)			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer Out (GB)	0.00026	South Central US	Storage	Resources			External
4/15/2010	Windows Azure Platform - All Services		North America	Data Transfer In (GB)	0.000002	South Central US	Storage	MODIS Source Data Products			External
4/15/2010	Windows Azure Storage			Storage (GB/month)	0.105552	South Central US	Storage	Resources			
4/15/2010	Windows Azure Storage			Storage Transactions (in 10,000s)	0.2866	South Central US	Storage	Resources			
4/15/2010	Windows Azure Storage			Storage (GB/month)	155.0917	South Central US	Storage	MODIS Source Data Products			
4/15/2010	Windows Azure Storage			Storage Transactions (in 10,000s)	4.846	South Central US	Storage	MODIS Source Data Products			
4/15/2010	Windows Azure Storage			Storage (GB/month)	14.84042	South Central US	Storage	Reduction Results			
4/15/2010	Windows Azure Storage			Storage Transactions (in 10,000s)	0.0006	South Central US	Storage	Reduction Results			
4/15/2010	Windows Azure Storage			Storage (GB/month)	92.50063	South Central US	Storage	Reprojection Results			
4/15/2010	Windows Azure Storage			Storage Transactions (in 10,000s)	0.0006	South Central US	Storage	Reprojection Results			

One day of ModisAzure billing



Cumulative MODISAzure billing ($39K)

100 instances @ $0.12 per hour = $288 per 24 hours

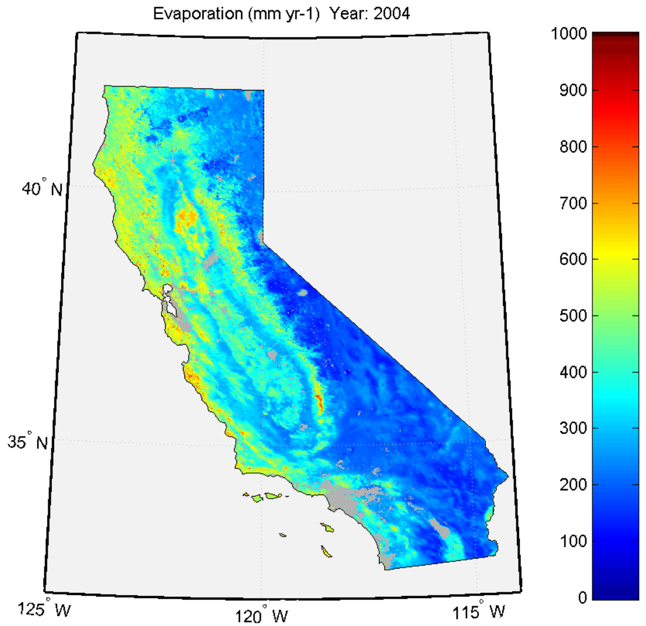
1 TB @ .15GB/mo = $150.

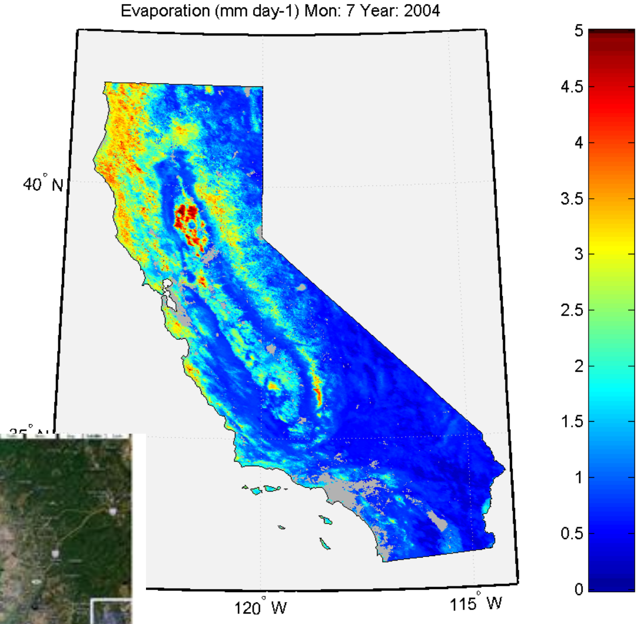
Summary

Lately it occurs to me
What a long strange trip it's been.
Grateful Dead

MODISAzure Learnings

- Putting all your eggs in the cloud basket means watching that basket
  - Cloud scale resources often mean you still manage small numbers of resources: 100 instances over 24 hours = $288 even if idle
  - Where is the long term archive for any results ?
- Azure is a rapidly moving target and unlike the Grid
  - Commercial cloud backed by large commercial development team
  - Current target applications are mid-range or smaller – MODISAzure is currently at the fringe
- Scaling up requires additional work as understanding even a 0.01% failure rate is time consuming
  - Bake in the faults for scaling and resilience
  - Bake in the catalog for end:end reconciliation of sources and results





TeraPixel Learnings

- DryadLINQ provides a powerful, elegant programming environment for large-scale data-parallel computing
- Trident Workflows reduce the barrier to modifying the flow while ensuring robust execution at scale

Tipping Points

- Handling the tsunami (even if it's just a small wave) of scientific data isn't quite computer science nor is it science.
  - Both can learn different things from joint work.
  - Computational science can (and may be the only way to) bridge the gap between the data glut and the scientist
- A few repeatable methodologies can generate that "perfect storm".
  - We (the computer science community) can seed that.
- If computing was free and people the only cost, what would we (the computer science community) advise?
  - Absolute performance is less important than time to science more important
  - Repeatability and provenance (by science definition) are key

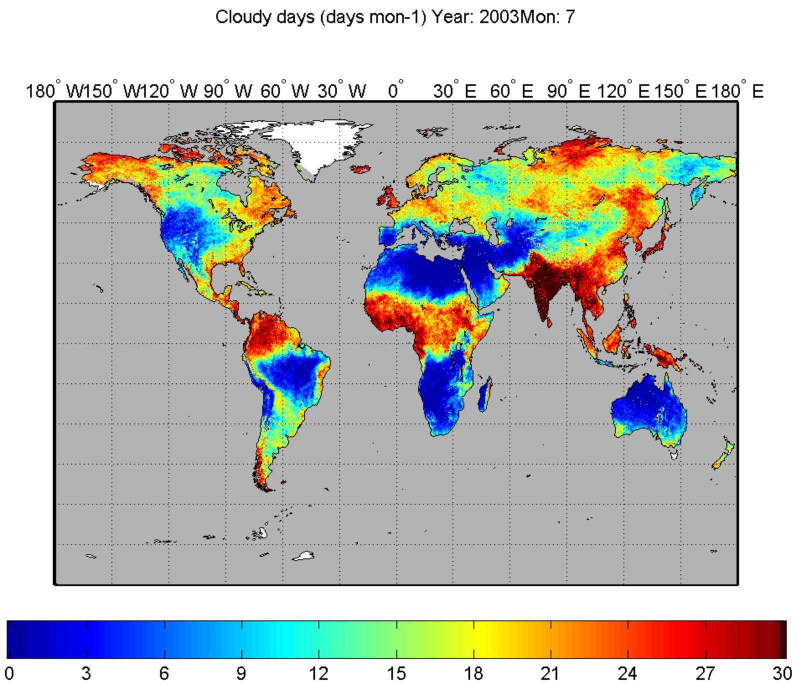
Acknowlegements

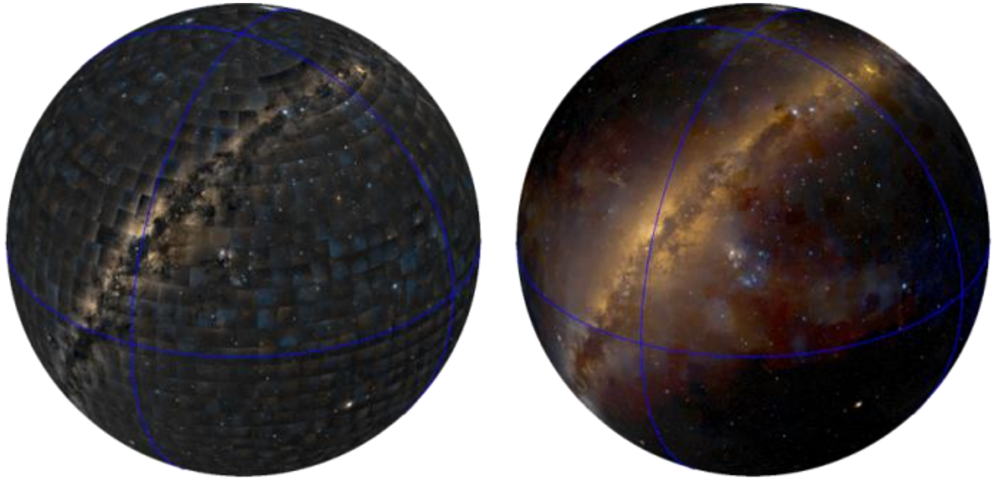
MODISAzure

- Scientists
  - Youngryel Ryu
  - Thomas Moran
  - Dennis Baldocchi
  - James Hunt
- Computer Scientists
  - Jie Li
  - You-Wei Cheah
  - Keith Jackson
  - Marty Humphrey
  - Deb Agarwal
  - Keith Beattie
- Others
  - The FLUXNET Collaboration
  - Roger Barga
  - Dan Fay
  - Dennis Gannon
  - David Heckerman
  - Tony Hey
  - Yogesh Simmhan

TeraPixel

- Dan Fay
- Jonathan Fay
- Dean Guo
- Christophe Poulain
- Hugues Hoppe
- Dennis Crain
- Mac Mason
- Brian McLean
- Michael Kazhdan





http://research.microsoft.com/terapixel

http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx